ANALYZING ALLELE SPECIFIC RNA EXPRESSION USING MIXTURE MODELS

Rong Lu[1], Ryan M. Smith[2], Michał Seweryn[1], Danxin Wang[2], Katherine Hartmann[2], Amy Webb[3], Wolfgang Sadee[2], and Grzegorz A Rempala[14]

[1]Division of Biostatistics, The Ohio State University [2]Center for Pharmacogenomics, The Ohio State University [3]Department of Biomedical Informatics, The Ohio State University [4]Mathematical Biosciences Institute, The Ohio State University

Introduction

Allele expression imbalance (AEI) or alternatively allele-specific gene expression (ASE) are used to describe the phenomenon when one parental copy of a given autosomal gene is preferentially expressed over the other in the corresponding RNA transcripts. The goal of AEI analysis is to separate the true signals (imbalanced expression due to biological mechanisms) from the noises (imbalanced expression due to instrumental variations and experimental biases in RNA-seq). Since imbalanced expression levels are used as the phenotype for identifying the responsible genetic variants, it is crucial to be able to get stable AEI analysis results without making unrealistic model assumptions.

Aims

The goal of this project is to develop appropriate statistical procedures for identifying AEI cases using RNA-seq read counts at heterozygous loci of different genes.

Methods

The term “folded Skellam” refers to the absolute value of the Skellam random variable. In the following model description, we denote the SNP allele reads from the paternal copy of a gene as *P* and that from the maternal copy as *M*.

Let *R* and *V* be the reference and variant reads respectively. Although the parental origin of reads is not available in our RNA-seq data, introducing the hidden pair *(P, M)* will help us in justifying the model for analyzing *(R, V)*.

In our current approach we only assume that Y: = P - M = Y_1 - Y_2 follows a Skellam mixture distribution with unknown fixed number of mixture components **K.**

$$\sum_{i=1}^{K}\left(\sum_{z=0}^{\min(p,m)}\pi_i \text{Poisson}(p-z\mid\lambda_{i,p})\text{Poisson}(m-z\mid\lambda_{i,m})f_{Z_i}(z)\right)$$

where $\{f_{Z_i}(z)\}_{i=1}^{K}$ is a set of unknown probability mass functions.

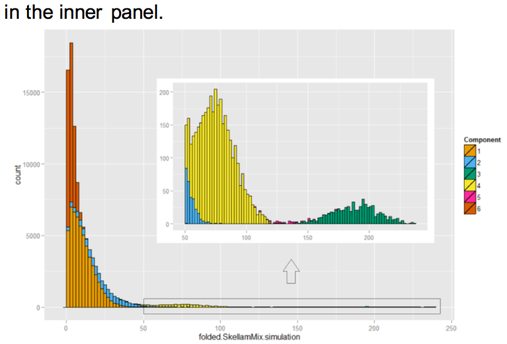
Methods cont.

Since we expect to have |R-V| = |P-M| it follows that |R-V| should have the same folded Skellam mixture distribution as |P-M| in our setting. Since the mean of the Skellam variable equals the difference of two corresponding Poisson means, testing the null hypothesis of no AEI signal within a mixture component is equivalent to testing whether the means of two independent Poisson variables are equal. That is, if the component i is a “no AEI signal” component, then under our model $\lambda_{i,p} = \lambda_{i,m} =: \lambda$ and we can estimate λ by the method of moments using the fact that $E(R-V)^2 = E(|R-V|)^2 = 2\lambda$.

Table 1 Poisson mixture model parameter estimates and SNPs classification results.

Mixture Component	Proportion	Poisson mean	No. of SNPs	No. of Genes
Comp.1	**0.030 (0.029, 0.031)**	**43.11 (42.54, 43.84)**	**18367**	**784**
Comp.2	0.0011 (0.0010, 0.0012)	152.37 (146.08, 166.13)	519	37
Comp.3	0.186 (0.182, 0.190)	20.34 (20.20, 20.49)	82963	3892
Comp.4	0.003 (0.0025, 0.0033)	108.14 (105.13, 115.60)	2073	89
Comp.5	0.0006 (0.0004, 0.0008)	201.01 (196.15, 209.71)	425	27
Comp.6	0.0073 (0.0069, 0.0077)	74.60 (72.56, 78.08)	5156	202
Comp.7	0.771 (0.769, 0.775)	7.82 (7.78, 7.85)	198889	11174

Figure 1 Histogram of the simulation from the fitted folded Skellam mixture (sample size=105). The two mixture components Mix1 and Mix6 which are closest to zero are considered the two no AEI signal components. The right tail (>50) with relatively smaller frequencies is enlarged and presented in the inner panel.



Results

The Poisson mixture model was fitted to the averaged total reads within tissue-specific genes (62326 tissue-specific genes in total, i.e. sample size=62326; overall log-likelihood=-216846; BIC=433836). Genes with the same rs number but from different brain region were considered as different tissue-specific genes.

Results cont.

We found the optimal number of mixture components to be 7, meaning that we could classify all SNPs into 7 “comparable” SNP groups. Most SNPs in the gene of our interest (SLC1A3) were classified into the mixture component Comp.1. The SNPs in Comp.1 were used to fit the folded Skellam mixture model.

In total 18367 SNPs were classified into the Poisson mixture component 1 and 10702 of them were in 3’ UTR of 531 genes. Fitting of the folded Skellam mixture model only used the 10702 SNPs in 3’ UTR.

Table 3 Folded Skellam mixture parameter estimates and results of LRTs for equal Poisson mean values.

Only SNPs on 3’ UTR and classified into Poisson mixture component 1 were used for fitting the folded Skellam mixture (overall log-likelihood=-34979; BIC=70117; sample-size=10702; $(\lambda\_(i,1), \lambda\_(i,2))$ is estimate of the ordered pair $(\lambda\_(i,P), \lambda\_(i,M))$.
NAs indicate insufficient sample sizes for LRTs.

Table 3 Folded Skellam mixture parameter estimates and results of LRTs for equal Poisson mean values.

Parameter	Mix1	Mix2	Mix3	Mix4	Mix5	Mix6
π_i	0.54 (0.54, 0.55)	0.1 (0.10, 0.11)	0.0065 (0.0064, 0.0066)	0.037 (0.036, 0.038)	0.0003 (0.0003, 0.00035)	0.3 (0.3, 0.31)
$\lambda_{i,1}$	65.7 (65.4, 66.5)	83.8 (82.6, 84.2)	268 (263.3, 269.4)	92.7 (91.4, 93.1)	214.8 (212.2, 216.3)	4.81 (4.75, 4.84)
$\lambda_{i,2}$	69.2 (69.2, 70.2)	106 (105, 107)	80.3 (79.9, 81.5)	166.0 (165.9, 169.1)	78.1 (77.0, 78.5)	5.39 (5.29, 5.40)
L_0	-17852	-2074	NA	-650	NA	-7860
L_1	-17864	-1967		-522		-8233
p-value	1	<0.00001		<0.00001		1
No. of SNPs	5459	482	3	130	2	4626
No. of Genes	471	165	3	72	2	407

Conclusion

By applying the folded Skellam mixture model to the RNA-seq data from human autopsy brain tissues, we find that 16% 3’UTR SNPs within a group of 531 ”comparable” genes show AEI. This result is consistent with the findings of other AEI studies in the literature.

Acknowledgements

This work was supported by the National Institute of General Medical Sciences (U01GM092655), the US National Science Foundation (DMS-1318886), and the US National Cancer Institute (R01-CA152158). This work was also supported in part by an allocation of computing time from the Ohio Supercomputer Center.

THE OHIO STATE UNIVERSITY